Published in final edited form as:

Stat Methods Med Res. 2017 December ; 26(6): 2821–2831. doi:10.1177/0962280215610928.

Unified variable selection in semi-parametric models

William Terry¹, Hongmei Zhang², Arnab Maity³, Hasan Arshad⁴, and Wilfried Karmaus²

¹Department of Mathematical Sciences, University of Memphis, Memphis, USA

²Division of Epidemiology, Biostatistics, and Environmental Health, School of Public Health, University of Memphis, Memphis, USA

³Department of Statistics, North Carolina State University, Raleigh, USA

⁴Allergy and Clinical Immunology, University of Southampton, Southampton, UK. The David Hide Asthma and Allergy Research Centre, Isle of Wight, UK

Abstract

We propose a Bayesian variable selection method in semi-parametric models with applications to genetic and epigenetic data (e.g., single nucleotide polymorphisms and DNA methylation, respectively). The data are individually standardized to reduce heterogeneity and facilitate simultaneous selection of categorical (single nucleotide polymorphisms) and continuous (DNA methylation) variables. The Gaussian reproducing kernel is applied to the transformed data to evaluate joint effect of the variables, which may include complex interactions between, e.g., single nucleotide polymorphisms and DNA methylation. Indicator variables are introduced to the model for the purpose of variable selection. The method is demonstrated and evaluated using simulations under different scenarios. We apply the method to identify informative DNA methylation sites and single nucleotide polymorphisms in a set of genes based on their joint effect on allergic sensitization. The selected single nucleotide polymorphisms and methylation sites have the potential to serve as early markers for allergy prediction, and consequently benefit medical and clinical research to prevent allergy before its manifestation.

Keywords

Bayesian methods; Gaussian kernel; non-linear effects; transformation; reproducing kernel; variable selection; single nucleotide polymorphisms; DNA methylation

1 Introduction

Motivated by research into the mechanisms and functionality of DNA that shows variations in DNA expression are due to more than just sequence variations,¹ this article discusses the use of methylation levels at various binding sites or CpG sites of DNA sequences as a

Reprints and permissions: sagepub.co.uk/journalsPermissions.nav

Corresponding author: Hongmei Zhang, Division of Epidemiology, Biostatistics, and Environmental Health, School of Public Health, University of Memphis, Memphis, TN 38152, USA. hzhang6@memphis.edu.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

potential predictor for susceptibility to allergens. Past research^{2,3} has focused on the effect of single nucleotide polymorphisms (SNPs) only or the effect of DNA methylation only. It is now considered that SNPs and DNA methylation work in concert to manifest a disease status.⁴⁻⁶ However, there is no method available with the ability to assess the contribution of complex interactions between CpG sites and between SNPs and CpG sites. To this end, the models and procedures developed in this article attempt to incorporate both SNPs and methylation at CpG sites as predictors in a single, unified model which has the potential to predict susceptibility to a disease or condition such as allergic reactions.

In order to incorporate potentially complex interactions of unknown linear or non-linear form between SNPs and CPG sites, semi-parametric models built upon reproducing kernels are considered.² Usually, an IBS kernel is applied to SNPs (categorical variables),^{3,7} and a Gaussian kernel is chosen for DNA methylation at a CpG site (representing continuous variables), but each of the two does not take a mixture of continuous and categorical variables, unless a mixture of kernels is implemented.⁸ In this article, we attempt to solve these problems using the ideas commonly applied in regression analyses. As seen in regression analysis,⁹ we can incorporate discrete variables into a regression model aptly suited for continuous variables. The same idea can be extended to reproducing kernels. Thus, instead of complex two kernels or mixture of kernels, we will only use the Gaussian kernels. However, the scale difference between categorical variables and continuous variables can cause severe bias in variable selection. This will be eased by variable standardization, a strategy commonly taken in regressions analyses.

To give the ability of variable selection to the semi-parametric model, we introduce a variable δ ,² which takes on values of 0 or 1 for non-inclusion and inclusion, respectively, for each candidate predictor. It is worth noting that our modeling approach allows for a large number of predictor variables with potentially vastly different scales and distributions.

2 Methods

Let $Y_{n \times 1}$ be a vector of response values for n subjects, $g_{n \times v}$ be a matrix of v potential predictor variables, and $X_{n \times p_0}$ be a matrix of p_0 known predictor variables with $\beta_{p_0 \times 1}$ being their linear effects. A model describing the association of X with Y and the effect of g on Y is assumed as

$$Y = X\beta + h(g) + \varepsilon, \quad (1)$$

where ε is random error, and $h(g)$ is a function of g which evaluates the effect of a set of variables g on Y . Variables g can be continuous or categorical. The function $h(g)$ may take linear or non-linear forms and represents the effect of possibly complex interactions. In this work, $h(\cdot)$ is defined using a semi-definite kernel function $K(\cdot)$. By the Mercer's theorem,^{10,11} under some regularity conditions, the kernel function $K(\cdot)$ specifies a unique function space H spanned by a particular set of orthogonal basis functions. Following the Mercer's theorem, any function $h(\cdot)$ in the function space H can be represented as a linear

combination of reproducing kernels,^{11–13} $h(g_i) = \sum K(g_i, g_k) \alpha_k = K_i' \alpha$; $i, k = 1, \dots, n$, where α is a vector of unknown parameters, and $K(g_i, g_k)$ is the Gaussian kernel defined as $K(g_i, g_k) = \exp(-\|g_i - g_k\|^2 / \rho)$, where ρ is the regularization parameter. The matrix formed by $K(g_i, g_k)$ is denoted as $K(\rho)$ representing a function of ρ . Note that $K(g_i, g_k)$ gives a measure of distance between observation i and k , and thus its definition applies to both continuous and categorical variables. For instance, assume we have one single nucleotide polymorphism (SNP) of interest. The distance can be 0 (completely the same SNP), 1 (one allele in common), or 2 (no alleles in common), which gives $K(g_i, g_k) = 1, \exp(-1/\rho)$, or $\exp(-4/\rho)$.

To select variables from g , we introduce an indicator variable $\delta = \{\delta_m, m = 1 \dots v\}$ into the kernel matrix with $\delta_m = 1$ denoting the inclusion of variable m and 0 otherwise. For instance, with $v=3$, $\delta = \{1, 1, 0\}$; the first two variables are selected. Accordingly, we update the notation of kernel matrix as $K(\rho, \delta)$ with its (i, k) -th entry defined as $K(g_i, g_k) = \exp(-\sum_m (\delta_m (g_{i,m} - g_{k,m}))^2 / \rho)$. If variable m is excluded, then it will not appear in any entry of the kernel matrix. The idea of using indicator variables for the inclusion or exclusion of a covariate has been applied in previous studies.^{3,14,15}

To infer the parameters, we take the fully Bayesian approach. In the next two sections, we define the prior distributions² and discuss posterior distribution calculations.

2.1 prior distributions

We start from the specification of prior distributions in the function $h(\cdot)$. The prior distribution of $h(\cdot)$ is adopted from Liu et al.¹⁶ and originated from a maximum likelihood estimator of $h(\cdot)$. A scaled penalized likelihood function for $h(\cdot)$ and coefficients β built upon model (1) is defined as¹⁶

$$J(h, \beta) = -\frac{1}{2} \sum_{i=1}^n (y_i - x_i^T \beta - h(g_i))^2 - \frac{1}{2} \lambda \|h(\cdot)\|_{H_K}^2. \quad (2)$$

It can be shown that the solution of $h(\cdot)$ by maximizing equation (2) is the same as that by assuming a prior distribution $h \sim N(0, \tau K)$.¹⁶ Due to the linear relationship between $h(\cdot)$ and α given $K(\cdot)$, taking the prior for $h(\cdot)$ above is equivalent to taking the prior distribution of α as $\alpha \sim N(0, \tau K^{-1})$. The variance component, τ , is assumed to follow the inverse gamma distribution, $\tau \sim \text{Inv} - \text{Gam}(a_\tau, b_\tau)$. The hyper-parameters a_τ and b_τ are known and small. We take the prior distribution for the variable inclusion parameter δ as Bernoulli, i.e., $\delta_l \sim \text{Ber}(1, q_l)$, choosing $q_l = q = 0.5$, $l = 1, \dots, v$ assuming no prior knowledge regarding variable inclusion is available. We now present the prior distributions for other parameters, specifically, the variance component σ^2 in ε and the coefficients β . We choose non-informative or vague priors. The prior distribution for σ^2 is assumed to be $\sigma^2 \sim \text{Inv} - \text{Gam}(a_{\sigma^2}, b_{\sigma^2})$ with known and small values for hyper-parameters a_{σ^2} and b_{σ^2} . In our simulations, we discuss the sensitivity to the variable selection results with respect to the choices of different prior distributions. The prior distribution for β is set as $\beta \sim N(0, \sigma_\beta^2 I)$,

where σ_β^2 is known and large and I denotes the identity matrix. Under different parameter spaces in terms of variables included in the model, different choices of ρ can produce the same likelihood. Thus, it is not applicable to estimate ρ while parameter spaces change. For now, we assume ρ is known and discuss its determination in Section 2.3.

2.2 Posterior distributions

Denote by $\Theta = \{h, \beta, \sigma^2, \tau, \delta\}$ a collection of parameters. The joint posterior distribution of Θ is, up to a normalizing constant

$$p(\Theta|Y, \rho) \propto p(Y|h, \beta, \sigma^2, \delta, \rho)p(h|\tau, \delta, \rho)p(\beta)p(\sigma^2)p(\tau)p(\delta|\rho). \quad (3)$$

To draw posterior inferences on Θ , we use the method of Markov chain Monte Carlo (MCMC) simulations, in particular, the Gibbs Sampler to sequentially draw samples from each parameter's full conditional posterior distribution. In the following $(\cdot)$ indicates all other parameters and the data whenever applicable. The conditional posterior of h is multivariate normal, $h(\cdot) \sim N(\Sigma_h(Y - X\beta)/\sigma^2, \Sigma_h)$, where $\Sigma_h = \{\tau^{-1}K^{-1} + (\sigma^2 I)^{-1}\}^{-1}$. For the coefficients β , their conditional posterior distribution is multivariate normal as well, $\beta(\cdot) \sim$

$$N(\Sigma_\beta X'(Y - h)/\sigma^2, \Sigma_\beta), \text{ where } \Sigma_\beta = \{X'X/\sigma^2 + I(\sigma_\beta^2)^{-1}\}^{-1}.$$

The conditional posterior distribution for τ is inverse-gamma, $\tau(\cdot) \sim \text{Inv-Gam}(n/2 + a_\tau, (h' \{K(\rho, \delta)\}^{-1}h + 2b_\tau)/2)$. A similar conditional posterior distribution for σ^2 can be derived, $\sigma^2(\cdot) \sim \text{Inv-Gam}(n/2 + a_{\sigma^2}, \{(Y - X\beta - h)'(Y - X\beta - h) + 2b_{\sigma^2}\}/2)$. Finally, parameter δ 's conditional posterior distribution is

$$Pr(\delta_l|(\cdot)) \propto \exp\{-(Y - X\beta - h)'(Y - X\beta - h)/(2\sigma^2) - [h' \{K(\rho, \delta)\}^{-1}h]/(2\tau)\} |K(\rho, \delta)|^{-1/2} q_l^{\delta_l} (1 - q_l)^{1 - \delta_l}. \quad (4)$$

Setting the right side of equation (4) as $cPr(\delta_l)$ with $a = cPr(\delta_l = 0|(\cdot))$ and $b = cPr(\delta_l = 1|(\cdot))$, we conclude a Bernoulli distribution with parameter $b/(b + a)$ as the conditional posterior for δ_l , $l = 1, \dots, v$.²

2.1.1 Determination of important variables—We estimate the posterior probabilities (posterior mean of δ) of including a variable in the model. To determine which variables should be kept, we apply the concept in scree plot to the posterior probabilities, calculated as the percentage of times that a variable is selected among a certain number of uncorrelated MCMC samples. Scree plots are often used in principal component analysis to determine the number of components, where a sharp decrease in eigenvalues indicates less importance of the rest of the components. Analogously, in our application of scree plots, a sharp decrease in probabilities indicates less importance of the remaining variables. Variables identified by

this rule are treated as the most important variables. In addition, a reference probability 0.50 can be used as well to identify a group of possibly important variables. This reference probability represents the expected frequency we randomly select a variable.

2.3 Re-scaling and estimation of ρ

Variables in different scales can potentially cause bias in inferences. To ease this problem, we standardize all the independent variables. Here we focus on the standardization of g . Let $\bar{g}_l$ and s_l denote the sample mean and standard deviation for predictor g_l respectively. Then it is known that the standardized variable g_l^* is defined as

$$g_l^* = \frac{g_l - \bar{g}_l}{s_l}, \quad (5)$$

which leads to the updated Gaussian kernel

$$K(g_{v1}^*, g_{v2}^*) = \exp\{-\|g_{v1}^* - g_{v2}^*\|^2 / \rho\}. \quad (6)$$

By standardizing, all variables are in the same scale. The functionality of the Gaussian kernel, however, is invariant to this transformation in that it still evaluates the distance between subjects. Standardizing all variables, although conveniently scaling variables of different magnitudes, has the drawback for ordinal variables of yielding only a few different values say (0,1,2). However, it is important to note that using ordinal variables in this statistical method as approximated by continuous distributions does not seem to bias inferences.^{17,18}

Turning to the determination of ρ , note that $\text{Var}(g_l^*) \rightarrow 1$ as $n \rightarrow \infty$. Linking this to the kernel calculations, we have $\text{Var}(g_{l1}^* - g_{l2}^*) \rightarrow 2$ assuming no correlations between the predictors. The stabilization in the predictors due to standardization motivated us to set ρ as $\rho=2$ and leave the task to τ to regularize the behavior of $h(\cdot)$. Our simulations discussed later support this choice and demonstrates its effectiveness and robustness.

We also estimate ρ by use of the Bayesian method based on the full model. In this case, we assume that unimportant variables do not contribute to the outcome variable of interest. We choose the prior distribution of ρ as *Inv-Gamma*(2, 2), which gives the prior mean of ρ as 2. This prior is chosen after observing the property of $\text{Var}(g_{l1}^* - g_{l2}^*)$. It is straightforward to assess that the conditional posterior of ρ is not in a standard form. We thus implement the Metropolis-Hastings algorithm in the Gibbs sampler to draw posterior samples of ρ . A jumping function for ρ is selected as a log-normal distribution with location parameter $\rho^{(t)}$ and variance parameter V , where $\rho^{(t)}$ is the ρ sampled at current iteration t , and V is chosen to achieve sampling efficiency. After ρ is estimated, it is then fixed through the variable selection process.

3 Results

3.1 Simulations

In total, 500 Monte Carlo (MC) replicates each with sample size n chosen from (25,50,100) and the number of predictor variables p chosen from (25,50) are simulated. Half of the predictor variables are generated from a gamma distribution with shape parameter 3 and scale parameter 0.5, $Gan(3, 0.5)$, and the other half are from a discrete uniform distribution, taking values (0,1,2). This data generating distribution represents a skewed distribution in each continuous component of g . This setting produces data analogous to data in our motivating example, logit-transformed DNA methylation (continuous measure) and SNPs (discrete measure), respectively. The random error component, ε , is generated from a $N(0, 0.5^2)$ with $\sigma^2 = 0.5^2$. We consider two different associations between variables g and Y , linear and non-linear associations,

$$\begin{aligned} \text{Model 1 (Linear)} & : E(Y_i | g_{snpi1}, g_{cpqi1}) = 1.5g_{snpi1} + 5g_{cpqi1} + g_{snpi1} \times g_{cpqi1} \\ \text{Model 2 (Non-linear)} & : E(Y_i | g_{snpi1}, g_{cpqi1}) = 5\cos(0.5 \times g_{snpi1} \times g_{cpqi1}), \end{aligned} \quad (7)$$

where, in model 2, only interaction effect is present.

For each simulation setting, to determine ρ via the Bayesian method, median of posterior draws of ρ in 10,000 MCMC iterations after 5000 burn-in is used as an estimate of ρ . For variable selections after ρ determined, we run three chains for purpose of convergence assessment with each chain having 10,000 iterations. The inferences presented in the article are from one chain based on 5000 iterations after 5000 burn-in iterations. An illustration of MCMC sequence convergence for parameters σ^2 and τ is given in Figures 1 and 2. For each model, we record three summary statistics: percentage of correct selection (CS) of the true model out of 500 MC replicates, along with average sensitivity (Sens) and specificity (Spec) and their standard deviations. Under the context of variable selection, sensitivity refers to the probability of selecting the truly important variables, and specificity refers to the probability of not selecting the truly unimportant variables.

The variable selection results for Models 1 and 2 are summarized in Table 1, where ρ is estimated based on the full model as discussed earlier. When the number of predictors is relatively small, e.g., $v=25$, the method can select the correct variables with high sensitivity and specificity (as well as high correctness rates), especially for Model 1 (linear models), e.g., with a sample size of 25, all the statistics (percentage of correct selection, sensitivity, and specificity) are perfect indicating a potential of handling the large p small n problem. In addition, for the non-linear model (Model 2), an improvement in the selection for each scenario is clearly observed as the sample size increases; with $v=25$ variables, the percentage of correct selection increases from 84.80% to 99.80%. When the number of variables is larger, e.g., $v=50$, the proposed method does not perform well, although an improving pattern of the three statistics is clearly observed in all the settings, e.g., the average sensitivity increased from 51.40% to 65.50% in Model 1. We also included results by setting $\rho=2$ (Table 2). Better selections are observed under this setting. Overall, the three statistics

improved significantly when the number of variables is 50 compared to taking ρ as the estimated quantity, especially when the sample size is 100.

3.2 Sensitivity analyses and comparisons

The above findings indicated a possibility of variable selection sensitivity with respect to the choice of ρ . Furthermore, the prior distributions of the variance components parameters τ and σ^2 are set at inverse gamma with small scale and shape parameters (both 0.001 in our above simulations); it is worth assessing the sensitivity with respect to this direction as well. To accomplish this, we use the 500 MC replicates each with sample size of 25 for both models that have 25 variables.

3.2.1 Sensitivity with respect to ρ —To assess the sensitivity of ρ , we consider different ρ s ranging from 0.5 to 10 when selecting the variables and compare the variable selection summary statistics from each ρ . The summary statistics are displayed in Table 3 and sorted by percentages of correct selections in descending order. As seen in the table, the selection results are robust as long as the choice of ρ is not far from its estimated values (in both cases, $\hat{\rho} = 1.29$).

It is also observed that if ρ is too low, e.g., $\rho = 0.5$, sensitivity is impacted more compared to specificity, and if it is too large, e.g., $\rho = 10$, then low specificity is more likely; in both cases, the percentages of correct selection are lower compared to other choices of ρ , especially in non-linear models.

3.2.2 Sensitivity with respect to prior distribution of τ and σ^2 —To assess whether different prior distributions of τ and σ^2 will produce substantial changes to the three statistics, percentages of correct selection (CS), sensitivity (Sens), and specificity (Spec), we considered 4 additional prior distributions. The first three prior distributions are still inverse gamma but with different shape and scale parameters, *Inv – Gam*(0.01, 0.01), *Inv – Gam*(0.1, 0.1), and *Inv – Gamma*(0.5, 0.5). The distribution *Inv – Gamma*(0.5, 0.5) was suggested by Kass and Wasserman (1995)¹⁹ and termed as “unit-information prior”. This prior distribution, approximately centered at 1, allows a greater influence from the data instead of the prior. The fourth prior distribution is a uniform distribution from 0 to 100, *Uniform*(0, 100), a noninformative but proper prior suggested in Gelman (2006).²⁰ We apply each of the four prior distributions to the 500 MC replicates and record the three statistics (percentage of correct selection, sensitivity and specificity) corresponding to each prior setting. In all these simulations, we set $\rho=2$ after observing the more stable variable selection results with this setting. Overall, compared to the results when the prior distribution is *Inv – Gam*(0.001, 0.001), the four additional prior distributions produced comparable results (Table 4), especially for sensitivity and specificity.

3.2.3 Comparison with existing methods—To demonstrate the advantage of the proposed method, we compare our approach with two existing methods. The first method is chosen from established variable selection approaches in linear models.

We take the recently proposed adaptive LASSO as a representative.²¹ For adaptive LASSO and almost all variable selection methods in linear models, the sample size needs to be larger

than the number of candidate predictors unless we presume the number of variables selected is smaller than sample size. Because of this, we considered the situation such that the number of variables is smaller than sample size, and thus only main effects are considered. The second method is suggested by Zhang et al.,² which has the ability to select variables with complex associations and these variables are with the same type, e.g., either continuous or categorical but not both. As noted earlier, the motivation of standardization is to scale variables of different magnitudes with the hope of reducing selection bias. To illustrate the need of standardization, we generated continuous variables from $Gam(10, 1)$, which produces data with much larger variation compared to $Gam(3, 0.5)$ used in our previous simulations. Since our goal is to assess the need of standardization, information loss should be avoided. Due to the periodicity of function $\cos(\cdot)$ in model 2, we lowered the frequency from 0.5 to 0.1 so that information passed to $\cos(\cdot)$ is comparable to that when generating from $Gam(3, 0.5)$. The prior distributions implemented in the second method are the same as those in Section 3.1 and we set $\rho=2$.

The method of adaptive LASSO gives lower percentages of correct selections (Table 5) compared to the proposed approach, which is expected due to the restriction of this type of variable selection approach.

However, the sensitivity and specificity are not impacted substantially. Coupled this observation with the low correct selection rates, the adaptive LASSO method is likely to over select variables but the number of over selected variables tends to be small. For the second method, we first performed variable selections separately for continuous and categorical variables as proposed in Zhang et al.² The percentages of correct selections are all lower than those in Tables 1 and 2 (results not shown). We next put both types of variables in the model but without standardization and performed variable selections. When the variation in the continuous variables is comparable to that in the categorical variables, standardizing the variables gives slightly better or comparable selection statistics (the first part of Table 6). However, when variations in the variables are not homogeneous (in our case, the variation in the continuous variables is much higher than that in the categorical variables), standardizing the variables substantially improves the selection statistics (the second part of Table 6).

3.3 Epigenetic application

We apply the proposed method to select informative CpG sites and SNPs potentially associated with wheal sizes resulting from skin prick tests for a wide array of allergens such as house dust mite or tree pollen. The subjects included in this study were from the Isle of Wight cohort.²² DNA methylation, SNPs, and wheal sizes are available for 202 women in the cohort. The sum of the wheal sizes for each skin prick tests is used as the response for each subject. The candidate predictors are 7 aptly chosen SNPs and 14 aptly chosen methylation sites on the GATA3 gene for a total of 21 predictor variables. These CpG sites and SNPs are selected due to their potential associations with atopy (wheal size at least 3 mm), an outcome related to asthma.²³

In total, 10,000 iterations are run and the posterior probabilities of δ for each predictor variable are recorded after a burn-in of 5000 iterations. Based on our simulations, we set

$\rho=2$. The posterior probabilities of including each candidate predictor are plotted in descending order (Figure 3). It is not surprising that the posterior probabilities are all larger than 0.5, since candidate SNPs and methylation sites are potentially informative, as noted above. To determine the final selection of variables, we use the strategy of scree plot noted earlier. Out of the 14 CpGs and 7 SNPs, 2 CpG sites (cg22770911 and cg00463367) and 2 SNPs (rs422628 and rs434645) are excluded (Tables 7 and 8). To further assess the sensibility of the proposed method, we considered all possible linear regression models that only include main effects of the 21 variables (CpGs and SNPs; in total, $\sim 2 \times 10^6$ models), and used Bayesian Information Criterion (BIC) to select the best model. The main effect model with three variables including CpG sites cg14327531, cg01255894, and SNP rs434645 give the best BIC, of which the two CpG sites were also selected by the proposed method with high posterior probabilities. SNP rs434645 was not selected by the new method, but its posterior probability was on the boundary of being selected. The other CpGs and SNPs uniquely selected by the proposed method are likely to have non-linear or complex interactions, which may be difficult to identify in linear models. Note that all the candidate CpG sites are either in the body or at 5' UTR, and half of the candidate SNPs are in introns. All these regions are non-coding regions and thus do not directly influence gene transcription. This potentially supports the idea of complex interactions between the CpG sites and SNPs, which indirectly influenced expression of genes and further wheal sizes.

4 Conclusion

The proposed method is an extension from a method for selecting one type of variables, either continuous or categorical but not both.² It has the ability to select both continuous and categorical variables. The selection is built into Gaussian kernels, which are commonly used for continuous variables and can be sensitive to different scales. We proposed to use standardized values in the kernel with both types of variables included, which eliminated the scale problem.

Simulations have demonstrated that the proposed method is effective in variable selection and robust with respect to different data generation schemes. As seen in the simulations, the proposed method has the potential to deal with a large number of candidate predictors, and allows the number of selected variables larger than sample size. Furthermore, it is ready to be extended to other models, e.g., probit regressions and survival analyses.

Most variable selection models rely on creating fully parametric models, linear or non-linear. For instance, the variable selection methods ALASSO or LASSO, which utilize approximations with linear basis functions, effectively select variables for linear or approximately linear models but perform poorly in nonlinear, periodic functions as seen in the earlier study.² Other approaches for selection of models and variables which rely on basis functions such as Fourier series would require pre-specification of a period length.

Recall that setting the tuning parameter ρ as $\rho=2$ following standardization provided better variable selection results. As discussed earlier, estimating $h(\cdot)$ was controlled by regularization parameters τ and ρ . After standardizing the data, the variations between different predictors were reduced, which potentially reduced the burden of ρ and τ . Fixing ρ

at $\rho=2$ and only using τ to regularize the behavior of $h(\cdot)$ in this case seems to work reasonably well, but a further investigation on this matter is certainly deserved.

Turning to the estimation of ρ , there is a possibility that different predictors may have different features. In this case, the corresponding regularization parameter ρ may be different for different features. One possible future work is to incorporate cluster analyses into the process of variable selection via clustering the variables and then inferring ρ for each cluster of variables. Nevertheless, the method proposed in this article is a starting point for selecting predictors in semi-parametric models built upon reproducing kernels from a mixture of continuous and categorical variables.

Acknowledgments

The authors thank Dr. Meredith Ray and Kranthi Guthikonda for their help and support.

Funding

This project was supported by NIH grants R01AI091905 (W Karmaus), R21AI099367 (H Zhang), and R00ES017744 (A Maity).

References

- Grundberg E, Meduri E, Sandling JK, et al. Global analysis of DNA methylation variation in adipose tissue from twins reveals links to disease-associated variants in distal regulatory elements. *Am J Hum Genet.* 2013; 93:876–890. [PubMed: 24183450]
- Zhang H, Maity A, Arshad H, et al. Variable selection in semi-parametric models. *Stat Methods Med Res.* Epub ahead of print 28 August 2013.
- He H, Zhang H, Maity A, et al. Power of a reproducing kernel-based method for testing the joint effect of a set of single-nucleotide polymorphisms. *Genetica.* 2012; 140:421–427. [PubMed: 23180006]
- Zhang H, Tong X, Holloway JW, et al. The interplay of DNA methylation over time with th2 pathway genetic variants on asthma risk and temporal asthma transition. *Clin Epigenet.* 2014; 6:8.
- Soto-Ramírez N, Arshad SH, Holloway JW, et al. The interaction of genetic variants and DNA methylation of the interleukin-4 receptor gene increase the risk of asthma at age 18 years. *Clinical Epigenetics.* 2013; 5:1. [PubMed: 23286427]
- Karmaus W, Ziyab AH, Everson T, et al. Epigenetic mechanisms and models in the origins of asthma. *Curr Opin Allergy Clin Immunol.* 2013; 13:63. [PubMed: 23242116]
- Wu MC, Kraft P, Epstein MP, et al. Powerful SNP-set analysis for case-control genome-wide association studies. *Am J Hum Genet.* 2010; 86:929–942. [PubMed: 20560208]
- Racine J, Li Q. Nonparametric estimation of regression functions with both categorical and continuous data. *J Econometr.* 2004; 119:99–130.
- Kutner, MH., Nachtsheim, C., Neter, J. *Applied linear regression models.* Chicago, IL: McGraw-Hill/Irwin; 2004.
- Mercer J. Functions of positive and negative type, and their connection with the theory of integral equations. *Philos Trans R Soc Lond Series A: Containing Papers Math Phys Char.* 1909:415–446.
- Cristianini, N., Shawe-Taylor, J. *An introduction to support vector machines and other kernel-based learning methods.* Cambridge, UK: Cambridge University Press; 2000.
- Kimeldorf GS, Wahba G. A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Ann Math Stat.* 1970:495–502.
- O'sullivan F, Yandell BS, Raynor WJ Jr. Automatic smoothing of regression functions in generalized linear models. *J Am Stat Assoc.* 1986; 81:96–103.
- George EI, McCulloch RE. Variable selection via Gibbs sampling. *J Am Stat Assoc.* 1993; 88:881–889.

15. Smith M, Kohn R. Nonparametric regression using Bayesian variable selection. *J Econometr.* 1996; 75:317–343.
16. Liu D, Lin X, Ghosh D. Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. *Biometrics.* 2007; 63:1079–1088. [PubMed: 18078480]
17. Zumbo BD, Zimmerman DW. Is the selection of statistical methods governed by level of measurement? *Can Psychol/Psychologie Canadienne.* 1993; 34:390.
18. Johnson DR, Creech JC. Ordinal measures in multiple indicator models: a simulation study of categorization error. *Am Soc Rev.* 1983:398–407.
19. Kass RE, Wasserman L. A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J Am Stat Assoc.* 1995; 90:928–934.
20. Gelman A, et al. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal.* 2006; 1:515–534.
21. Zou H. The adaptive lasso and its oracle properties. *J Am Stat Assoc.* 2006; 101:1418–1429.
22. Arshad SH, Hide DW. Effect of environmental factors on the development of allergic disorders in infancy. *J Allergy Clin Immun.* 1992; 90:235–241. [PubMed: 1500628]
23. Guthikonda K, Zhang H, Nolan VG, et al. Oral contraceptives modify the effect of gata3 polymorphisms on the risk of asthma at the age of 18 years via DNA methylation. *Clin Epigenet.* 2014; 6:17.

Figure 1.
 σ^2 Convergence.

Figure 2.
 τ Convergence.

Figure 3.
Posterior probabilities of δ_v .

Table 1

Simulation results for variable selections (ρ estimated).

Model	ρ	Samp.	# vars. (categ., cont.)	% CS	Sens	Spec
1 (lin.)	1.29	25	25 (12, 13)	100	100 (0)	100 (0)
	1.29		50 (25, 25)	0	51.40 (34.94)	49.41 (6.87)
	1.31	50	25 (12, 13)	100	100 (0)	100 (0)
	1.29		50 (25, 25)	1.00	49.40 (35.38)	50.59 (8.76)
	1.34	100	25 (12, 13)	100	100 (0)	100 (0)
2 (non-lin.)	1.30		50 (25, 25)	28.40	65.50 (36.70)	63.97 (23.54)
	1.29	25	25 (12, 13)	84.80	94.90 (18.43)	96.06 (12.75)
	1.29		50 (25, 25)	0	47.30 (36.33)	49.97 (7.04)
	1.31	50	25 (12, 13)	99.80	99.90 (2.24)	99.97 (0.58)
	1.29		50 (25, 25)	0.40	50 (36.36)	50.51 (8.07)
	1.33	100	25 (12, 13)	100	100 (0)	100 (0)
	1.29		50 (25, 25)	8.00	57.20 (35.92)	53.76 (15.14)

CS: correct selection; Sens: sensitivity; Spec: specificity.

Table 2

Simulation results for variable selection ($\rho = 2$).

Model	ρ	Samp.	# vars. (categ., cont.)	% CS	Sens	Spec
1 (lin.)	2	25	25 (13, 12)	100	100 (0)	100 (0)
		50	50 (25, 25)	0.4	49.10 (37.37)	50.05 (7.74)
	50	25	25 (13, 12)	100	100 (0)	100 (0)
		50	50 (25, 25)	21.02	62.50 (37.03)	61.15 (21.19)
2 (non-lin.)	2	25	25 (13, 12)	100	100 (0)	100 (0)
		50	50 (25, 25)	99.40	99.70 (4.00)	99.72 (3.60)
	50	25	25 (13, 12)	88.00	98.40 (9.88)	98.40 (7.38)
		50	50 (25, 25)	0	49.80 (35.53)	50.05 (6.85)
	100	25	25 (13, 12)	99.80	100 (0)	99.99 (0.20)
		50	50 (25, 25)	1.80	52.50 (34.51)	50.82 (9.88)
		100	25 (13, 12)	100	100 (0)	100 (0)
			50 (25, 25)	74.00	86.80 (29.11)	86.85 (22.45)

CS: correct selection; Sens: sensitivity; Spec: specificity.

Table 3

Simulation results on variable selection sensitivity with different choices of ρ .

Model	ρ	Samp.	# vars. (categ., cont.)	% CS	Sens	Spec
1 (lin.)	2	25	25 (13, 12)	100	100 (0)	100 (0)
	1.29			100	100 (0)	100 (0)
	1			99.80	99.90 (2.24)	100 (0)
	3			99.80	100 (0)	99.99 (0.20)
	5			98.20	100 (0)	99.92 (0.58)
	10			91.20	100 (0)	99.60 (1.28)
2 (non-lin.)	0.5			90.40	95.80 (15.58)	97.48 (11.14)
	2	25	25 (13,12)	88.00	98.40 (9.88)	98.40 (7.38)
	1.29			84.80	94.90 (18.43)	96.06 (12.75)
	3			82.60	98.90 (8.00)	98.78 (3.53)
	1			77.20	91.70 (22.74)	92.82 (16.98)
	5			65.80	99.10 (7.37)	97.35 (5.60)
	0.5			42.80	72.50 (34.36)	77.46 (24.59)
	10			29.40	98.90 (7.80)	92.35 (7.48)

Note: The results are sorted by descending %CS for each model.

CS: correct selection; Sens: sensitivity; Spec: specificity.

Table 4

Simulation results on variable selection sensitivity with the choice of prior distributions for τ and σ^2 .

Model	Priors of τ, σ^2	% CS	Sens	Spec
1 (lin.)	$Inv - \text{Gam}(0.001, 0.001)$	100	100 (0)	100 (0)
	$Inv - \text{Gam}(0.01, 0.01)$	100	100 (0)	100 (0)
	$Inv - \text{Gam}(0.1, 0.1)$	100	100 (0)	100 (0)
	$Inv - \text{Gam}(0.5, 0.5)$	100	100 (0)	100 (0)
	$Uniform(0, 100)$	100	100 (0)	100 (0)
2 (non-lin.)	$Inv - \text{Gam}(0.001, 0.001)$	88.00	98.40 (9.88)	98.40 (7.38)
	$Inv - \text{Gam}(0.01, 0.01)$	90.00	98.10 (10.56)	98.42 (7.75)
	$Inv - \text{Gam}(0.1, 0.1)$	94.20	98.70 (9.67)	99.04 (5.99)
	$Inv - \text{Gam}(0.5, 0.5)$	95.40	98.20 (11.27)	98.06 (9.98)
	$Uniform(0, 100)$	93.80	98.00 (11.23)	98.18 (8.98)

CS: correct selection; Sens: sensitivity; Spec: specificity.

Table 5

Simulation results for variable selections using the method of adaptive LASSO.

Model	Samp.	# vars. (categ., cont.)	% CS	Sens	Spec
1 (lin.)	50	25 (13,12)	66.20	100 (0)	96.14 (7.95)
	100	25 (13,12)	77.60	100 (0)	98.58 (3.11)
	100	50 (25,25)	75.60	100 (0)	98.93 (2.61)
2 (non-lin.)	50	25 (13,12)	55.20	99.40 (5.45)	95.67 (7.10)
	100	25 (13,12)	73.60	100 (0)	98.19 (3.65)
	100	50 (25,25)	68.80	100 (0)	98.66 (2.68)

CS: correct selection; Sens: sensitivity; Spec: specificity.

Table 6

Simulation results for variable selection sensitivity using the method by Zhang et al.²

Model	% CS	Sens	Spec	% CS	Sens	Spec
Continuous variables generated from $Gam(3,0.5)$ Variables standardized						
1 (lin.)	100	100 (0)	100 (0)	99.20	100 (0)	99.96 (0.39)
2 (non-lin.)	88.00	98.40 (9.88)	98.40 (7.38)	82.20	98.20 (11.70)	98.51 (5.10)
Continuous variables generated from $Gam(10, 1)$ Variables standardized						
1 (lin.)	100	100 (0)	100 (0)	83.00	91.40 (23.60)	93.47 (16.76)
2 (non-lin.)	96.60	99.20 (6.28)	99.82 (0.98)	35.60	69.20 (28.15)	88.71 (17.46)

CS: correct selection, Sens: sensitivity, Spec: specificity.

Table 7

Selection of CpG sites in the *GATA3* gene on wheal size at 18 years.

CpG site	location	position	selected	frequency	Index
cg18599069	5'UTR	8096991	✓	0.62	8
cg10008757	5'UTR	8097183	✓	0.64	9
cg14327531	5'UTR	8097331	✓	0.76	10
cg17124583	Body	8097641	✓	0.64	11
cg11430077	Body	8099018	✓	0.72	12
cg01255894	Body	8099218	✓	0.70	13
cg10089865	Body	8100286	✓	0.66	14
cg22770911	Body	8101307		0.50	15
cg04492228	Body	8101513	✓	0.68	16
cg17489908	Body	8101566	✓	0.70	17
cg03669298	Body	8102210	✓	0.66	18
cg00463367	Body	8103673		0.50	19
cg04213746	Body	8106003	✓	0.62	20
cg27409129	Body	8111731	✓	0.56	21

Table 8
Selection of single nucleotide polymorphisms (SNPs) in the *GATA3* gene on wheal size at 18 years.

SNP	position	location	selected	frequency	Index
rs1269486	8096199	Promoter	✓	0.62	2
rs3802604	8102272	Intron	✓	0.58	3
rs3824662	8104208	Intron	✓	0.62	4
rs422628	8111409	Intron		0.52	6
rs406103	8111621	Intron (boundary)	✓	0.62	5
rs434645	8121451	3' UTR		0.52	7
rs12412241	8127139	Downstream	✓	0.72	1